

HMM-BASED GLISSANDO DETECTION FOR RECORDINGS OF CHINESE BAMBOO FLUTE

Changhong Wang¹, Emmanouil Benetos¹, Xiaojie Meng², Elaine Chew¹

¹Centre for Digital Music, Queen Mary University of London, UK

{changhong.wang, emmanouil.benetos, elaine.chew}@qmul.ac.uk

²Department of Chinese Music, China Conservatory of Music, China

mxj.yd@foxmail.com

ABSTRACT

Playing techniques such as ornamentations and articulation effects constitute important aspects of music performance. However, their computational analysis is still at an early stage due to a lack of instrument diversity, established methodologies and informative data. Focusing on the Chinese bamboo flute, we introduce a two-stage glissando detection system based on hidden Markov models (HMMs) with Gaussian mixtures. A rule-based segmentation process extracts glissando candidates that are consecutive note changes in the same direction. Glissandi are then identified by two HMMs. The study uses a newly created dataset of Chinese bamboo flute recordings, including both isolated glissandi and real-world pieces. The results, based on both frame- and segment-based evaluation for ascending and descending glissandi respectively, confirm the feasibility of the proposed method for glissando detection. Better detection performance of ascending glissandi over descending ones is obtained due to their more regular patterns. Inaccurate pitch estimation forms a main obstacle for successful fully-automated glissando detection. The dataset and method can be used for performance analysis.

1. INTRODUCTION

Computational analysis of expressive patterns in music signals plays an important role in music information research. For instrumental music, these expressive patterns are frequently the result of playing techniques. Automated analysis of playing techniques can benefit automatic music transcription [1], computer-aided music pedagogy [2], instrument classification [3, 4], and performance analysis [5]. However, computational analysis of playing techniques is still in its early stages, lacking instrument diversity, established methodologies, and informative data.

Most existing work on computational analysis of playing techniques focuses on Western instruments such as guitar [6–8], violin [9–11], piano [12], and drums [13, 14].

C. Wang is funded by the China Scholarship Council (CSC). E. Benetos is supported by a UK RAEng Research Fellowship (RF/128).

Copyright: © 2019 Changhong Wang, Emmanouil Benetos, Xiaojie Meng, Elaine Chew. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Playing techniques in non-Western instruments, while similarly important, are often overlooked. Take for example, one of the world’s most ancient instruments, the Chinese bamboo flute (also known as the *Dizi* or *Zhudi*, thereafter referred to as CBF): many listeners are most often captivated by its unique timbre, which belies the twenty or more playing techniques invoked when performing on the instrument. To our knowledge, only Ayers [15, 16] has done some analysis of CBF playing techniques through synthesis. This work focused only on trills, tremolos and flutter-tongue. But many other techniques remain to be explored. For the case of other non-Western instruments, limited computational work can be found [5, 17].

For playing technique detection, methods adopted in the literature are typically frame-wise classifiers based on high dimensional feature inputs [6, 18], with little explanation of why the methods work. Support vector machines (SVMs) are the most frequently used class of methods. A series of electric bass guitar playing techniques was classified into plucking or expressive styles using SVMs in [6]; [10] applied it to distinguish five fundamental guitar playing techniques. A multimodal input using SVMs was used for analysing piano pedalling techniques in [12]. Su et al. [11] proposed new features as input to an SVM based on sparse modeling of magnitude and phase-derived spectra before classifying violin playing techniques. Other work used dynamic time warping [19], COSFIRE filters [20], spectrogram templates [21], and filter diagonalisation method [22] for analysis of playing techniques.

Datasets used in playing technique research consist of mainly playing techniques performed in isolation. Isolated techniques can vary greatly from the same techniques used in live performance. For ecological validity, we argue that playing techniques should be collected in context. A challenge of obtaining playing technique examples in real-world settings is that some techniques may be rare. Thus, it may be hard to find pieces covering a wide range of playing techniques and with sufficient repeated instances of these techniques to obtain a variety of samples for a specific technique.

To address these limitations, we use the CBF as our instrument of choice and glissando, a rarely explored audio gesture in the literature, as our starting point, aiming to build a systematic methodology for automatically analysing playing techniques. *Glissando*, here refers to a rapid slide up or down the musical scale [23], which is not

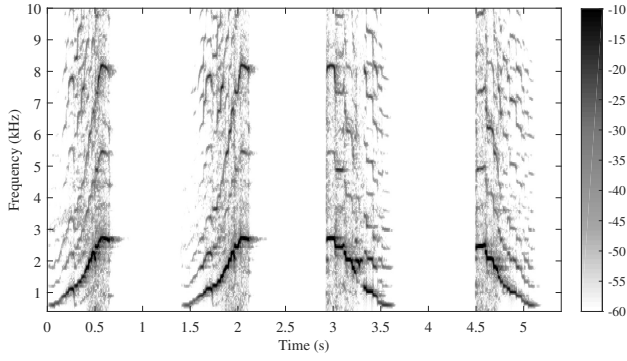


Figure 1. Spectrogram of two ascending and two descending glissando examples in Chinese bamboo flute music.

comparable to the one defined as a continuous slide from one note to another in [24]. Fig. 1 shows a spectrogram of a series of two ascending and two descending CBF glissandi. As can be seen, they exhibit a readily recognisable pattern, resembling rapid scale segments. Glissando detection in CBF playing is not straightforward: CBF glissandi are less regular than the stair-like glissando patterns in piano and guitar playing [18]. For the same glissando type, variations exist in the ways they are executed between different players, different pieces, and even different parts of the same piece. The main characteristic of glissando is the consecutive note change, which we claim can be captured by latent states of a hidden Markov model (HMM) [25, 26]. HMMs enable the decoding of note evolution while smoothing outlier variations within performed glissandi.

In this paper, we make a first attempt to the computational analysis of CBF glissandi. A new dataset including both isolated glissandi and real-world pieces is created and is being prepared for public release. Based on the analysis of ground truth statistics, we propose a two-stage detection system. A rule-based segmentation process first extracts glissando candidates that are consecutive note changes in the same direction. Different from traditional binary classification, the false positives obtained in the segmentation stage, which exhibit similar pitch evolution and duration as the ground truth, are used to train a non-glissando HMM (NG-HMM). A glissando HMM (G-HMM) is trained using all ground truth glissandi in the training set. Glissandi are then identified by two HMMs at test time.

2. DATASET

2.1 Dataset Information

The glissando analysis dataset, CBF-GlissDB, comprises recordings by ten expert CBF players from the China Conservatory of Music. All data is recorded in a professional recording studio using a Zoom H6 recorder at 44.1kHz/24-bits. Each player performs both isolated glissandi covering all notes on the CBF and one full-length piece—*Busy Delivering Harvest* 《扬鞭催马运粮忙》 or *Morning* 《早晨》. Players are grouped by flute type (C and G, the most representative types for Southern and Northern styles, respectively) and each player uses their own flute. Details of

recording length and number of glissandi in each group are shown in Table. 1.

Players	Flute	Isolated glissandi		Whole-piece recordings	
		Length (mins)	#glissandi [↑, ↓]	Piece, style	Length (mins) #glissandi [↑, ↓]
1-3	C	2.4	[58,47]	<i>Morning</i> , Southern	16.0 [24,0]
4-10	G	5.0	[117,112]	<i>Busy Delivering Harvest</i> , Northern	28.0 [23,106]

Table 1. Dataset information.

In order to assess the performance of the proposed glissando detection system independent of the performance of pitch estimation methods, pitch ground truth for all recordings is created. The fundamental frequency of each recording is first estimated using the pYIN algorithm [27] due to the strictly monophonic property of the recordings. All errors are then manually corrected by the first author using Sonic Visualiser¹. Both isolated and performed glissandi are annotated and verified by the players on the score. The final annotation is created by the first author after consulting with the players.

2.2 Dataset Statistics

To verify the intuition of the difference between isolated and performed glissandi, characteristic statistics of the ground truth are calculated. Fig. 2 shows two-dimensional histograms for four types of glissandi in CBF-GlissDB: ascending and descending isolated glissandi; and ascending and descending performed glissandi. As can be seen, performed glissandi have shorter durations than isolated glissandi, especially for descending glissandi, performed ones have almost half duration as isolated ones. Further analysis of note durations within each glissandi shows little difference among isolated glissandi while ascending performed glissandi have larger variation than descending performed ones. This may be attributed to the performers' tendency to lengthen the start or end note in an ascending performed glissando.

3. METHOD

To automatically detect glissando from real-world CBF recordings, we propose a two-stage detection system based on rule-based segmentation (Sec. 3.1) and HMM-based identification (Sec. 3.2).

3.1 Rule-based Segmentation

To obtain glissando candidates from the whole-piece recordings, we introduce a rule-based segmentation component using pitch with a 20ms hop size as input, as demonstrated in Fig. 3. The pitch is first smoothed to exclude noisy variations and quantised to the nearest notes in 12-tone equal temperament scale, resulting in 16 notes in the CBF tonal

¹ <https://www.sonicvisualiser.org>

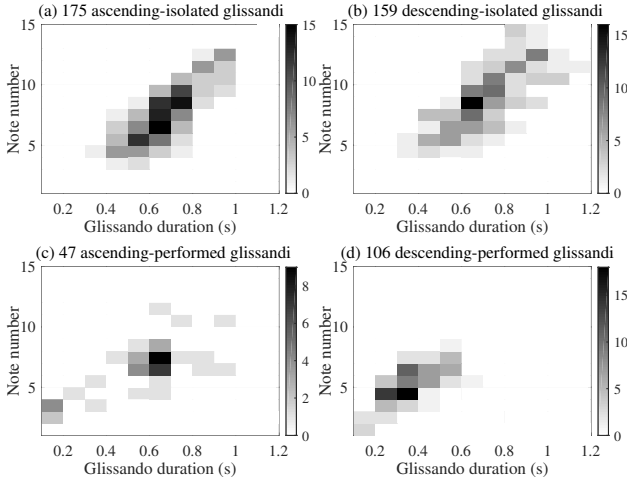


Figure 2. Duration and note number histograms for four glissando types.

range: G4-A6 for the C flute, and D5-E7 for the G flute (we assume that flute types are known for the current system). Frames with pitch less than 250Hz and waveform amplitude less than -20dB are marked as silence. The sign of note change is extracted to represent note change direction. Consecutive note changes in the same direction are then extracted as glissando candidates, which are further pruned by constraints on note numbers (at least 4 for both ascending and descending glissandi) and duration (at least 0.2s for ascending glissandi and 0.15s for descending glissandi based on the consultations with the professional players).

3.2 HMM-based Identification

3.2.1 Feature Extraction

Since all glissando candidates (extracted in the previous stage) share similar pitch evolution characteristics, the input to the HMMs must possess sufficient discriminative power to distinguish glissandi from non-glissandi. Considering the pitch discreteness and long duration of glissandi, we use a feature set consisting of both short-term (average pitch change, average intensity, average intensity change) and long-term (note number, note duration, note range) features [28, 29]. All features are statistics (mean and standard deviation) of pitch and intensity with variations on window and hop sizes. Hop size variations range from 10ms to 20ms at intervals of 2ms, while window sizes depend on the glissando direction.

(i) Short-term features:

To capture pitch and intensity change, the short-term window varies from 100 to 200ms at intervals of 20ms for the following three features.

– Average pitch change:

$$\Delta p_i = \frac{1}{w} \sum_{k=1}^w [p_i(k) - p_{i-1}(k)], \quad (1)$$

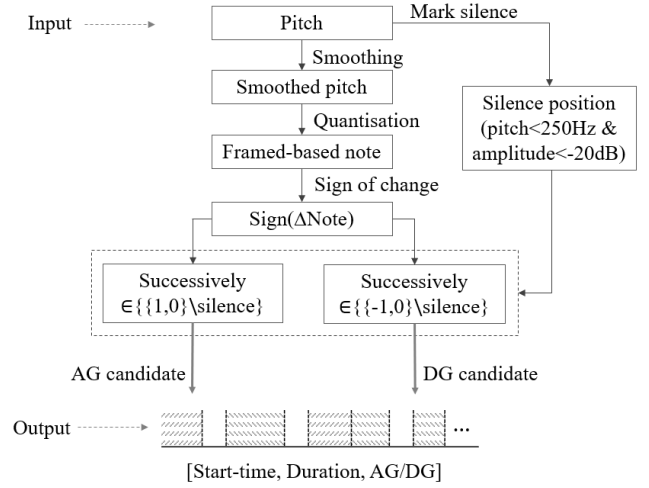


Figure 3. Diagram of rule-based segmentation (AG=ascending glissando; DG=descending glissando).

where $p_i(k)$ is the k -th pitch value within the window centered at the i -th time frame, and w is the window length.

– Average intensity (amplitude in dB scale) [7]:

$$I_i = \frac{1}{w} \sum_{k=1}^w [20 \cdot \log_{10} A_i(k)], \quad (2)$$

where $A_i(k)$ is the amplitude of the k -th sample within the window centered at the i -th time frame, and I_i is average intensity of this window.

– Average intensity change: $\Delta I_i = I_i - I_{i-1}$.

(ii) Long-term features:

To capture the discreteness of pitch evolution, note-level features with long windows are calculated. The window sizes vary from 200 to 400ms at intervals of 50ms for descending glissandi with shorter duration, and from 200 to 600ms at the same intervals for ascending glissandi which have longer duration. The calculation process for one ascending glissando example is shown in Fig. 4. With a 400ms window sliding forward, the number of notes N is 8 (one more than the number of peaks, highlighted by the red circles) and note range (note change between start and end notes) R equals 7. Note durations D , which refer to the intervals between two note change peaks, are {80,40,60,40,40,60}ms.

3.2.2 HMM-based Identification

As shown in Fig. 5, two HMMs with Gaussian mixture emissions are trained on the training set, with k-means initialisation and iterative parametrisation by the Expectation-Maximization algorithm [30]. During the training process, model parameters—the number of HMM latent states, number of Gaussian mixture components, and window-hop sizes—are varied and the model with the best performance on

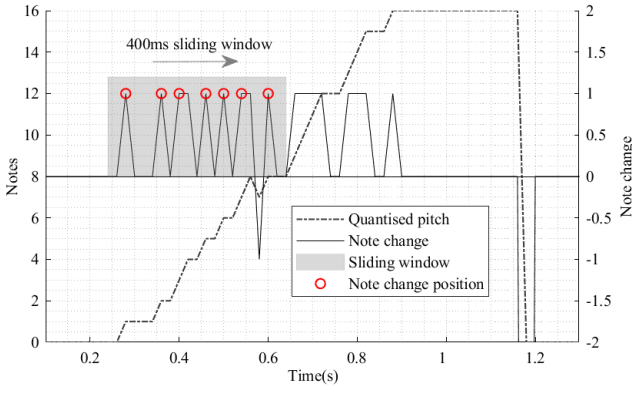


Figure 4. Long-term feature calculation process based on one example of an ascending glissando.

the validation set is chosen as the final one for testing. The emission used is a Gaussian mixture distribution [30]:

$$p(\mathbf{x}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m=1}^M \pi_m \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (3)$$

where \mathbf{x}_i is the observed feature vector of the i -th frame; π_m , $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the prior, mean and covariance of the m -th mixture component; and $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are the model parameters, each of which is an M -dimensional vector corresponding to π_m , $\boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_m$.

The CBF-GlissDB is subdivided into three subsets, namely, training (all isolated glissandi and 6 whole pieces), validation (2 whole pieces), and test (2 whole pieces). The segmentation stage is applied to whole-piece recordings in all three subsets, but to different ends. For the training set, segmentation serves the purpose of obtaining false positives that are then used to train a NG-HMM. In the validation and test stages, the extracted segments serve as candidates to be assigned glissando (G) or non-glissando (NG) labels by comparing the log-likelihood calculated by the two HMMs. Since the HMMs are applied directly to the candidate segments, the absolute position of glissandi in the pieces does not influence the result. The ten whole-piece recordings are randomly allocated to the training, validation, and test sets in a 6:2:2 ratio at the beginning of experiment. A five-fold cross-validation is then conducted.

4. EVALUATION

To investigate the influence of automatic pitch estimation on glissando detection, evaluation of both a semi-automated system (using the pitch ground truth as input) and a fully-automated system (using pitch automatically estimated by pYIN [27] as input) is carried out. Because glissando length ranges approximately from 200 to 1100ms, for each system, frame-based and segment-based evaluations are implemented. The frame size used in frame-based evaluation is 20ms. Segment-based evaluation compares detected glissandi and ground truth in short-time, non-overlapping segments [31]. A segment length of 100ms is adopted. True positives are segments which have overlaps with both

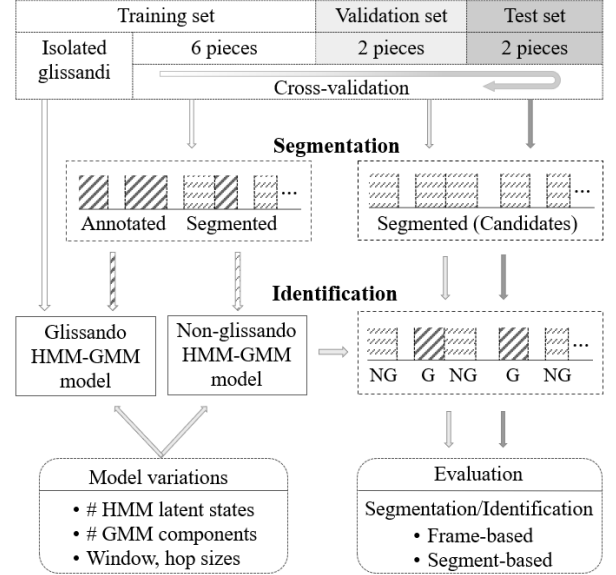


Figure 5. System diagram for glissando detection (G=glissando; NG=non-glissando).

ground truth and detected glissandi; false positives segments overlaps only with detected glissandi; and, false negatives intersect with ground truth only.

4.1 Semi-automated System Evaluation

Table 2 gives the precision, recall, and F-measure results for both ascending and descending glissandi in the semi-automated detection system. As can be seen, the segmentation stage performs a conservative selection of candidate segments with high recall and low precision. The large number of false positives obtained for NG-HMM training benefits the data balance in our system. The better identification performance of ascending glissandi over descending ones can be attributed to their more regular patterns. As can be seen, the identification F-measure increased by approximately 60% as compared to the segmentation F-measure, which verifies our intuition that consecutive pitch evolution can be captured by HMMs.

Stage	Glissando direction	Frame-based (%)			Segment-based (%)		
		\mathcal{P}	\mathcal{R}	\mathcal{F}	\mathcal{P}	\mathcal{R}	\mathcal{F}
Rule-based segmentation	Ascending	3.1	93.4	5.9	3.1	92.8	6.0
	Descending	4.9	83.1	9.0	5.1	86.9	9.9
HMM-based Identification	Ascending	73.4	75.4	73.4	72.0	74.0	72.0
	Descending	65.4	67.6	63.2	64.4	70.2	64.2

Table 2. Evaluation results of the semi-automated glissando detection system based on annotated pitch (\mathcal{P} =precision, \mathcal{R} =recall, \mathcal{F} =F-measure).

4.2 Fully-automated System Evaluation

After verifying the proposed glissando detection method independently, we then use the automatically estimated pit-

ch to evaluate the fully-automated glissando detection system. Due to the influence of breathing, some parts in the CBF recordings have high intensity but no detected pitch. Thus silence cannot be determined only by pitch presence, and we define silence bits as parts having both no pitch and intensity below -20dB. Correctly detected frames are the voiced parts with pitch intervals less than half a semitone between the ground truth and the detected pitch. Pitch estimation accuracy refers to the percentage of correctly detected frames over all voiced frames. Table 3 shows the estimated pitch result of both whole-piece recordings and ground truth glissando segments within these pieces. The poorer pitch estimation performance on glissando segments shows that pYIN works less well on rapid pitch evolution progressions.

Type	Whole pieces		Glissando segments	
	Southern	Northern	Ascending	Descending
Accuracy (%)	80.2	79.5	72.0	74.8

Table 3. Pitch estimation accuracy for whole-piece recordings and glissando segments.

The fully-automated glissando detection results are shown in Table 4. Considering the pitch evaluation shown above, it is reasonable to expect worse performance when using automatically estimated pitch as input. Pitch is a main discriminative feature in the proposed glissando detection system. The presence of undetected pitches or octave errors within glissandi hinders G-HMM to capture the consecutive note evolution. Thus false positives, which exhibit similar pitch evaluation as the ground truth glissandi and have higher pitch estimation accuracy, may be assigned with G labels. This is verified by the better identification performance on descending glissandi over ascending ones with lower pitch estimation result.

Stage	Glissando direction	Frame-based (%)			Segment-based (%)		
		\mathcal{P}	\mathcal{R}	\mathcal{F}	\mathcal{P}	\mathcal{R}	\mathcal{F}
Rule-based segmentation	Ascending	2.1	84.8	4.1	2.1	86.2	4.4
	Descending	3.3	67.3	5.9	3.6	75.0	7.1
HMM-based identification	Ascending	36.4	63.2	44.6	36.8	63.4	45.0
	Descending	58.2	48.4	50.4	58.0	51.8	52.6

Table 4. Evaluation results of the fully-automated glissando detection system based on estimated pitch.

5. CONCLUSIONS

In this paper, we have described a first attempt at computational analysis of CBF glissandi. HMMs are introduced to decode the consecutive note evolution within glissandi and a two-stage detection system is proposed. Using inputs based only on the statistics of two low-level features—pitch and intensity, frame- and segment-based F-measures of 73.4% and 72.0% for ascending glissandi, and 63.2%

and 64.2% for descending glissandi, are obtained in a semi-automated detection system, which confirms the feasibility of our method for glissando detection. The poorer performance of the fully-automated system may be attributed to the inaccuracy of pitch estimation since pitch is the main discriminative feature.

Future work will seek to implement other state-of-art pitch estimation methods (for example, CREPE [32]) to improve pitch detection accuracy prior to glissando detection. More informative features for glissando description may be explored. Alternative methods for glissando identification will be investigated, such as template-based detection, the spiral scattering transform [33], and deep learning, the latter including data augmentation of the collected audio samples. Plans are underway for expansion of the dataset. The analysis will also be expanded to other CBF playing techniques, with the aim to develop a systematic methodology for CBF playing technique detection.

6. REFERENCES

- [1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: challenges and future directions,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.
- [2] Y. Han and K. Lee, “Hierarchical approach to detect common mistakes of beginner flute players,” in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 77–82.
- [3] G. E. Hall, H. Ezzaidi, M. Bahoura, and C. Volat, “Classification of pizzicato and sustained articulations,” in *European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.
- [4] V. Lostanlen, J. Andén, and M. Lagrange, “Extended playing techniques: the next milestone in musical instrument recognition,” in *5th International Conference on Digital Libraries for Musicology*, 2018.
- [5] L. Yang, “Computational modelling and analysis of vibrato and portamento in expressive music performance,” Ph.D. dissertation, Queen Mary University of London, 2017.
- [6] J. Abeßer, H. Lukashevich, and G. Schuller, “Feature-based extraction of plucking and expression styles of the electric bass guitar,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 2290–2293.
- [7] J. Abeßer and G. Schuller, “Instrument-centered music transcription of solo bass guitar recordings,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 9, pp. 1741–1750, 2017.
- [8] L. Su, L. F. Yu, and Y. H. Yang, “Sparse cepstral and phase cdes for guitar playing technique classification,” in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 9–14.

- [9] I. Barbancho, C. Bandera, A. M. Barbancho, and L. J. Tardon, "Transcription and expressiveness detection system for violin music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 189–192.
- [10] S. H. Chen, S. H. Wu, Y. S. Lee, R. Lo, and J. C. Wang, "Hierarchical representation based on Bayesian non-parametric tree-structured mixture model for playing technique classification," in *Proceedings of the Thematic Workshops of ACM Multimedia*, 2017, pp. 537–543.
- [11] L. Su, H. M. Lin, and Y. H. Yang, "Sparse modeling of magnitude and phase-derived spectra for playing technique classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2122–2132, 2014.
- [12] B. Liang, G. Fazekas, A. McPherson, and M. Sandler, "Piano pedaller: a measurement system for classification and visualisation of piano pedalling techniques," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2017, pp. 325–329.
- [13] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," in *International Conference on Music and Artificial Intelligence (ICMAI)*, 2002, pp. 69–80.
- [14] M. Prockup, E. M. Schmidt, J. J. Scott, and Y. E. Kim, "Toward understanding expressive percussion through content based analysis," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 143–148.
- [15] L. Ayers, "Synthesizing trills for the Chinese dizi," in *International Computer Music Conference (ICMC)*. Singapore, 2003, pp. 227–30.
- [16] —, "Synthesizing timbre tremolos and flutter tonguing on wind instruments," in *International Computer Music Conference (ICMC)*. Miami, Florida, USA, 2004.
- [17] T. H. Özaslan, X. Serra, and J. L. Arcos, "Characterization of embellishments in ney performances of makam music in Turkey," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 13–18.
- [18] Y. P. Chen, L. Su, and Y. H. Yang, "Electric guitar playing technique detection in real-world recording based on F0 sequence pattern recognition," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 708–714.
- [19] S. Giraldo and R. Ramírez, "Performance to score sequence matching for automatic ornament detection in jazz music," in *International Conference of New Music Concepts (ICMNC)*, 2015.
- [20] A. Neocleous, G. Azzopardi, C. N. Schizas, and N. Petkov, "Filter-based approach for ornamentation detection and recognition in singing folk music," in *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2015, pp. 558–569.
- [21] J. Driedger, S. Balke, S. Ewert, and M. Müller, "Template-based vibrato analysis in music signals," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 239–245.
- [22] L. Yang, K. Rajab, and E. Chew, "The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation," *Journal of Mathematics and Music*, vol. 11, no. 1, pp. 42–60, 2017.
- [23] Merriam-Webster, *Webster's ninth new collegiate dictionary*. Merriam-Webster, 1983.
- [24] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1, 2018.
- [25] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [26] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.
- [27] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.
- [28] P. C. Li, Y. H. Y. L. Su, and A. W. Y. Su, "Analysis of expressive musical terms in violin using score-informed and expression-based audio features," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 809–815.
- [29] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project." technical report, IRCAM, Paris, France, Apr. 2004.
- [30] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT Press, 2013.
- [31] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.
- [32] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.
- [33] V. Lostanlen and S. Mallat, "Wavelet scattering on the pitch spiral," in *18th International Conference on Digital Audio Effects (DAFx)*, 2015.